University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2016

Analyses of transcriptome sequences reveal multiple ancient large-scale duplication events in the ancestor of Sphagnopsida (Bryophyta)

Devos, Nicolas ; Szövényi, Peter ; Weston, David J ; Rothfels, Carl J ; Johnson, Matthew G ; Shaw, A Jonathan

DOI: https://doi.org/10.1111/nph.13887

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-131191
Journal Article
Accepted Version

Originally published at:
Devos, Nicolas; Szövényi, Peter; Weston, David J; Rothfels, Carl J; Johnson, Matthew G; Shaw, A Jonathan (2016). Analyses of transcriptome sequences reveal multiple ancient large-scale duplication events in the ancestor of Sphagnopsida (Bryophyta). New Phytologist, 211(1):300-318.
DOI: https://doi.org/10.1111/nph.13887

New Phytologist

Analyses of transcriptome sequences reveal multiple ancient large-scale duplication events in the ancestor of Sphagnopsida (Bryophyta)

Journal:	*New Phytologist*
Manuscript ID	NPH-MS-2015-20654.R1
Manuscript Type:	MS - Regular Manuscript
Date Submitted by the Author:	n/a
Complete List of Authors:	Devos, Nicolas; Duke University, Biology; Szovenyi, Peter; University of Zurich, Institute of Systematic Botany; University of Zurich, Institute of Evolutionary Biology and Environmental Studies; ELTE, MTA ELTE-MTM Ecology Research Group; Quartier Sorge-Batiment Genopode, Swiss Institute of Bioinformatics Weston, David; Oak Ridge National Laboratory, Biosciences Division Rothfels, Carl; University of British Columbia, Zoology Johnson, Matthew; Chicago Botanic Garden, NA Shaw, A. Jonathan; Duke University, Department of Biology;
Key Words:	Whole genome duplication (WGD), Ks plot, transcriptome, peatmoss, Sphagnum, reconciliation, molecular dating, paleopolyploidy

SCHOLARONE™
Manuscripts

Analyses of transcriptome sequences reveal multiple ancient large-scale duplication events in the ancestor of Sphagnopsida (Bryophyta)

Nicolas Devos[1*], Péter Szövényi[2,3,4,5*], David J. Weston[6], Carl J. Rothfels[7], Matthew G. Johnson and A. Jonathan Shaw[1]

[1]Department of Biology, Duke University, Durham, North Carolina, USA 27708

[2]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland.

[3]Institute of Systematic Botany, University of Zurich, Zurich, Switzerland.

[4]Swiss Institute of Bioinformatics, Quartier Sorge-Batiment Genopode, Lausanne, Switzerland.

[5]MTA ELTE-MTM Ecology Research Group, ELTE, Biological Institute, Budapest, Hungary.

[6]Oak Ridge National Laboratory, Biosciences Division , Oak Ridge, Tennessee, USA 37831

[7]University Herbarium & Department of Integrative Biology, University of California, Berkeley, Berkeley, California, USA 24720

[8]Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022

*Shared first and corresponding authors

Authors for correspondence:
Nicolas Devos
nd28@duke.edu
Duke University, Department of Biology
Box 90338 Durham, NC 27708, USA
Tel: +1 (919)660 7298

For Peer Review

Péter Szövényi
peter.szoevenyi@uzh.ch
University of Zurich, Institute of Systematic Botany
Zollikerstr 107, Zurich, CH-8008, Switzerland
Tel: +41 (0) 634 84 18

Total word count: 6311
Introduction: 1076
Materials and Methods: 2529
Results: 1190
Discussion: 1349
Acknowledgements: 135
Number of figures:5
Number of tables:6
Supporting information: 1

For Peer Review

Summary

- The goal of this research is test whether there has been a whole genome duplication (WGD) in the ancestry of *Sphagnum* (peatmoss) or the class Sphagnopsida, and to determine if the timing of any such duplication(s) and patterns of paralog retention could help explain the rapid radiation and current ecological dominance of peatmosses.

- Illumina RNA-seq data were generated for nine taxa in Sphagnopsida (Bryophyta). Analyses of frequency plots for synonymous substitutions per synonymous site (K_s) between paralogous gene pairs and reconciliation of 578 gene trees were conducted to assess evidence of large-scale or genome-wide duplication events in each transcriptome.

- Both K_s frequency plots and gene tree-based analyses indicate multiple

duplication events in the history of the Sphagnopsida. The most recent WGD event predates divergence of *Sphagnum* from the two other genera of Sphagnopsida. Duplicate retention is highly variable across species, which might be best explained by local adaptation.

- Our analyses indicate that the last WGD could have been an important factor underlying the diversification of peatmosses and facilitated their rise to ecological dominance in peatlands. The timing of the duplication events and their significance in the evolutionary history of peat mosses is discussed.

Keywords: Whole genome duplication (WGD), K_s plot, reconciliation, molecular dating, transcriptome, paleopolyploidy, peatmoss, *Sphagnum*

Introduction

Recent evidence has revealed at least 50 independent whole genome duplication (WGD) events distributed across the angiosperm tree of life (Vision *et al.*, 2000; Bowers *et al.*, 2003; Blanc & Wolfe, 2004; Paterson *et al.*, 2004; Schlueter *et al.*, 2004; Van de Peer & Meyer, 2005; Cannon *et al.*, 2006; Cui *et al.*, 2006; Tuskan *et al.*, 2006; Jaillon *et al.*, 2007; Barker *et al.*, 2008, 2009; Lyons *et al.*, 2008; Ming *et al.*, 2008; Soltis *et al.*, 2009; Shi *et al.*, 2010; Van de Peer, 2011; Jiao *et al.*, 2012; McKain *et al.*, 2012; Tayale & Parisod, 2013; Amborella Genome Consortium, 2013). WGD can introduce new genetic variation for evolution to act upon (Ohno, 1970); nevertheless, its contributions to organismal diversification are still debated. Recent studies suggest that polyploid lineages have diversified at a slower pace than their diploid relatives, at least over relatively short time scales (Mayrose *et al.*, 2011, Arrigo & Baker, 2012; Escudero *et al.,* 2014; Scarpino *et al.*, 2014; Mayrose *et al.*, 2014). Conversely, others have argued that polyploidization is a major driver of diversification and that inferred lower diversification rates of polyploid species is probably a methodological artifact (Soltis *et al.*, 2009; Van de Peer *et al.,* 2009; Doyle, 2012; Schranz *et al.*, 2012; Amborella Genome Project, 2013; Cannon

For Peer Review

92 *et al.*, 2014; Soltis *et al.,* 2014). WGDs have also been suggested as a factor affecting
93 survival of plant lineages through the catastrophic Cretaceous-Tertiary extinction event
94 (Fawcett *et al.*, 2009; Vanneste *et al.*, 2014). In addition, WGDs may have contributed to
95 the evolution of form and complexity during early land-plant evolution through the
96 duplication of key regulatory genes (Rensing, 2014).
97
98 A feature of past whole-genome duplications is that only a relatively small proportion of
99 duplicate loci are retained (Ohta, 1987; Walsh, 1995; Vision *et al.*, 2000; Blanc & Wolfe
100 2004; Paterson *et al.*, 2004; Van de Peer & Meyer, 2005; Cui *et al.,* 2006; Jaillon *et al.*,
101 2007; Shi *et al.*, 2010; Jiao *et al.*, 2012; Amborella Genome Consortium 2013). It is
102 generally observed that paralog retention is non-random and is primarily driven by
103 natural selection (Freeling *et al.,* 2008; Freeling, 2009; Makino & McLysaght, 2010;
104 Birchler & Veitia 2011; Liu *et al.*, 2011; Barker *et al.*, 2012; Schnable *et al.*, 2012; Chen
105 *et al.*, 2013; Conant *et al.*, 2014). In spite of many years of research it is still unclear why
106 some gene duplicates are retained while others are lost or undergo divergent evolution.
107 Paralog retention could be explained by the dosage sensitivity of genes, sub- or
108 neofunctionalization, by gene-specific features, by lineage-specific evolutionary forces,
109 or by a combination of these factors (Birchler & Veitia, 2011; Barker *et al.*, 2012;
110 Carretero-Paulet & Fares, 2012; Jiang *et al.*, 2013; Conant *et al.*, 2014).
111
112 The number, timing, and evolutionary impact of WGD events in angiosperms has been
113 the focus of research for many years (see citations above), but relatively little is known in
114 this respect about other land plants, such as bryophytes. Data on the two model mosses,
115 *Physcomitrella patens* and *Ceratodon purpureus*, suggest that both genomes underwent
116 independent but chronologically coincident large-scale duplication events (Rensing *et al.*,
117 2007; Szövényi *et al.*, 2015) and that paralog retention in these genomes deviated from
118 the patterns found in angiosperms. In particular, genes involved in signal transduction
119 and transcriptional regulation appear *not* to have been preferentially retained in mosses in
120 contrast to genes involved in general metabolism that seem to have been preferentially
121 preserved at least in *P. patens*. Whether these observations are specific to *P. patens* and
122 *C. purpureus*, to mosses, or to all nonvascular land plants (i.e., liverworts, hornworts, and

For Peer Review

mosses) is currently unclear. Similarly, it is unknown when the WGDs inferred for *P. patens* and *C. purpureus* occurred relative to moss diversification. That is, can the same WGDs found in *P. patens* and *C. purpureus* be detected in other mosses? Finally, in contrast to the angiosperms, the association between WGD and diversification rates has not been investigated for the bryophytes.

Peatmosses (*Sphagnum*) dominate many wetland habitats, especially at northern latitudes. *Sphagnum* species are quintessential ecosystem engineers, not only dominating peatland habitats where they occur, but actually creating the environmental conditions that contribute to their dominance (Turetsky *et al.*, 2008). A time-calibrated phylogenetic reconstruction for the genus suggests that it underwent a rapid and relatively recent diversification that may have corresponded to periods of climatic cooling during and after the Miocene (Shaw *et al.*, 2010a). This timing corresponds to the earliest appearance of *Sphagnum*-dominated peatlands in the fossil record (Greb *et al*., 2006). If this is the case, the diversification of *Sphagnum* represents a striking radiation of a group into new habitats. With some 30% of terrestrial carbon currently bound in *Sphagnum*-dominated peatlands (Vasander *et al.*, 2006), this radiation was globally important in the evolution of earth's biogeochemistry.

The Sphagnopsida comprise one of four major speciose clades within the phylum Bryophyta and it is an early diverging group within mosses (Goffinet & Shaw, 2009; Chang & Graham, 2011; Wickett *et al.*, 2014). Based on phylogenetic analyses, Shaw *et al.* (2010b) recognized four genera in three families within the Sphagnopsida: *Ambuchanania* and *Eosphagnum* (Ambuchananiaceae), *Flatbergium* (Flatbergiaceae), and *Sphagnum* (Sphagnaceae). *Ambuchanania*, *Eosphagnum*, and *Flatbergium* currently comprise one species each, whereas *Sphagnum* includes 250 – 450 species. (We informally include a second species in *Flatbergium* based on unpublished molecular evidence). It is clear that *Ambuchanania*, *Flabergium*, and *Eosphagnum* are phylogenetically outside *Sphagnum*; in the following, we refer to these two groups as the non-*Sphagnum* and *Sphagnum* peatmosses, respectively (Fig. 1).

For Peer Review

The goals of this research were to test, using comparative transcriptome data, whether there is a signal of one or more ancestral genome duplications in the Sphagnopsida, and to determine if the timing of any such duplication(s) and patterns of paralog retention could help explain the rapid radiation and current ecological dominance of peatmosses. We used next-generation sequencing technology to sequence the transcriptomes of nine species of Sphagnopsida, and inferred large-scale duplications from the distribution of pairwise genetic divergence between duplicated genes expressed as the number of synonymous substitutions per synonymous sites (K_s). This method is a reliable means of inferring relatively recent large-scale duplications in the absence of genome-wide synteny information (Vanneste *et al.*, 2013). We then verified these findings by analyzing hundreds of gene trees using tree reconciliation in a phylogenomic framework to estimate the relative timing of the WGD. Finally, we used molecular dating to estimate the absolute timing of the duplication event.

Materials and Methods

Transcriptome data were newly generated for *Eosphagnum inretortum*, *Flatbergium sericeum*, *Sphagnum (Flatbergium) novo-caledoniae*, *Sphagnum cribrosum*, *S. subsecundum* and *S. fallax* (Fig. 1). Hereafter in this paper we informally refer to the species currently known as *Sphagnum novo-caledoniae* as "Flatbergium novo-caledoniae" to reflect phylogenetic relationships (Shaw, unpublished); this nomenclatural transfer to *Flatbergium* has not been formalized yet. Transcriptome data for *S. recurvum*, *S. lescurii* (Matasci *et al.*, 2014) and S. *palustre*, generated for the 1kp project, were also included in this study. All nine species other than *S. palustre* have haploid gametophytes. *Sphagnum palustre* is a relatively recently formed allopolyploid (Karlin *et al.*, 2010).

RNA extraction and sequencing

Total RNA was isolated from fresh gametophytic tissue using a Spectrum™ Plant Total RNA kit (Sigma) following the manufacturer's instructions. Extracted RNA was eluted

For Peer Review

into a final volume of 100 µL RNase-free water. RNA quantity was estimated using Qubit (Life Technologies, Carlsbad, California, USA) and RNA quality was estimated using an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, California, USA). For each taxon, one RNA extract with a RIN score greater than 8.0 was used for library preparation and sequencing. RNA-seq paired-end libraries were constructed with average fragment lengths of 300 base pairs (bp). Libraries for *E. inretortum*, *F. sericeum* and "F. novo-caledoniae*"* were indexed, pooled, and sequenced on a single lane of an Illumina HiSeq 2000 sequencer flow cell (Illumina, San Diego, California, USA). The *S. cribrosum* library was part of a multiplexed (2 samples pooled) run on one lane of Illumina HiSeq 2000. All Illumina sequencing was done at The Duke University Genome Sequencing & Analysis Core Resource and generated 100bp paired-end sequences. For *S. fallax* we used the mRNA to synthesize a double-stranded cDNA library with the cDNA Rapid Library Preparation protocol provided by Roche and sequenced it using two full 454 FLX plates at Oakridge National Laboratory. Specimen voucher information and accession numbers for the transcriptome data are provided in Table 1.

Transcriptome assembly and ORF predictions

Illumina data form *S. cribrosum*, *S. subsecundum*, "F. novo-caledoniae", *E. inretortum* and *F. sericeum* were assembled using Trinity_r20131110 (Grabherr *et al.*, 2011; Haas *et al.*, 2013) whereas data generated in the 1Kp project for *S. recurvum*, *S. palustre* and *S. lescurii* were processed by SOAPdenovo-Trans 1.03 (Xie *et al.*, 2014) using default parameters. The 454 data for *S. fallax* was filtered, trimmed and assembled using Newbler v2.5.3 (454 Life Sciences) with default options. To confirm that any putative signal of whole genome duplication would be robust to the algorithm used to assemble the short reads into transcripts, the transcriptome of "Flatbergium novo-caledoniae" was assembled with both assemblers and the consistency of duplication analysis assessed. Prior to assembly, the 3-prime ends of each read were trimmed based on quality score using the FASTQ Quality Trimmer (FASTX-Toolkit, http://hannonlab.cshl.edu/fastx_toolkit/) with a quality score threshold of 25 (Sanger/Illumina 1.9 encoding). Illumina sequencing adapters, when present, were

For Peer Review

clipped off using FASTQ/A Clipper (FASTX-Toolkit) and only paired-end reads for which both sequences were longer than 40bp after trimming and clipping were kept and assembled into transcripts.

Transcriptome assemblies were blasted against the Uniprot database (The UniProt Consortium, 2014) using BLASTX v2.2.25 (Altschul *et al.*, 1997). To remove transcripts assembled from contaminating organisms, BLASTX outputs were filtered to discard transcripts with top hits to non-land plants. Blast outputs were further filtered to remove all transcripts for which top hits had an e-value > 10^{-6}. This step also ensured that only high-confidence protein translations were kept with homologs in other Viridiplantae. We note that in this step some peat moss-specific sequences may have been discarded. Nevertheless, in the absence of a peat moss genome this strategy ensured reliable filtering of the data while minimizing the number of false positives.

Open reading frame (ORF) predictions for all transcripts that passed these filtering steps were made using the "transcripts_to_best_scoring_ORFs" perl script provided with the Trinity package. For each transcript, the longest, best-scoring predicted ORF with a length of at least 150 amino acids were used for further downstream analyses.

Analyses of potential whole genome duplication(s)

Redundancy filtering

After assembly and contamination filtering, the transcriptomes were still unusually large compared to other moss transcriptomes (*P. patens;* Zimmer *et al.*, 2013, *C. purpureus;* Szövényi *et al.*, 2015, *Syntrichia caninervis*; Gao *et al.*, 2014), suggesting incomplete assembly. To reduce artifactual redundancy we clustered predicted coding sequences of transcripts (CDS) with CD-HIT (Fu *et al.*, 2012). When generating clusters we used a similarity threshold of 98% and required that 90% of the length of the putatively redundant sequence had to align to the longest sequence. We kept only the longest putative CDS sequence per cluster.

Paralog identification and gene clustering

For each taxon, we first took protein predictions of the filtered data sets and clustered

For Peer Review

transcripts into groups containing two or more paralog sequences using the algorithm provided in BlastClust (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html). We used single linkage clustering with the following threshold values: e-value$\leq 10^{-10}$ and 30% similarity over 75% of the length of the protein sequence (Ponting & Russel, 2002). We then generated all pairwise alignments of paralogs within each cluster. Peptide sequences belonging to a paralog pair were aligned globally using MUSCLE v3.7 (Edgar, 2004). Nucleotide sequences were then forced onto the corresponding amino acid sequence alignments using PAL2NAL v13 (Suyama *et al.*, 2006). Finally, pairwise K_s values were calculated for each nucleotide alignment using K_aK_s-Calculator (Zhang *et al.*, 2006) under the YN model of sequence evolution (Yang & Nielsen, 2000).

The number of possible pairs of paralogs within a paralog cluster with more than two members is greater than the number of duplication events. Therefore, we generated simple phylogenetic trees for each paralog cluster by single-linkage clustering using the matrix of pairwise K_s values for each species separately. We then calculated K_s values for each node of the trees (nodal K_s values) as the simple average of pairwise K_s values. Each node of the paralog clusters represents one duplication event and thus nodal K_s values are the appropriate chronological representation of the duplication process. We used nodal K_s frequency plots to detect potential whole-genome duplications in the nine Sphagnopsida species.

Peaks produced by paleopolyploidy are expected to be approximately Gaussian (Blanc & Wolfe, 2004; Schlueter *et al.*, 2004). Therefore, multivariate normal components were fitted to the resulting K_s frequency distributions using the mixture model test implemented in the program EMMIX (McLachlan *et al.*, 1999) to identify peaks corresponding to putative large-scale duplications. The mixed distributions were modeled with one to ten components and the Expectation Maximization (EM) algorithm was repeated 100 times with random starting values. The optimal number of components in the mixture model was identified using the Bayesian information criterion (BIC) as the optimality criterion.

Relative Timing of WGD

For Peer Review

To identify the timing of putative WGD events relative to the split between non-*Sphagnum* peatmoss species and *Sphagnum* sensu stricto species, we compared the K_s distribution of non-Sphagnum and *Sphagnum* paralogs with the K_s distribution of non-Sphagnum and *Sphagnum* orthologs. An ortholog distribution with a mode smaller than its paralog counterpart indicates that the *Sphagnum*/non-Sphagnum divergence event occurred after the putative WGD, while the reverse indicates that independent WGD events in the two lineages occurred after their divergence from each other. Sequences from two accessions were defined as orthologs if they were each other's best hit, they aligned over >150 amino acids, and had a similarity of at least 30% in a BLASTP search. K_s values for each orthologous pair were then calculated using the procedure described above and plotted together with K_s values for paralogs. This procedure was done only for three species-pair comparisons ("F. novo-caledoniae*" – S. lescurii*, *F. sericeum– S. recurvum*, and *E. inretortum – S. palustre*). We selected species spanning the taxonomic diversity of *Sphagnum* and non-*Sphagnum* peatmosses. *S. lescurii*, *S. recurvum* and *S. palustre* are representatives of the three large sections (monophyletic clades) of *Sphagnum* peatmosses while we used all three species of the non-*Sphagnum* peatmosses.

Tree-based hypothesis testing of large-scale duplications

K_s-distribution based estimates of large-scale duplication cannot account for heterogeneity in evolutionary rates among lineages/genes, suffer from mutational saturation, and preferentially concentrate paralogs at more recent nodes (Vanneste et al., 2013). Therefore, we supplemented our K_s-based analysis with a phylogenomic approach utilizing gene tree-species tree reconciliation to provide relative estimates for the timing of duplication. To build gene families we used all ORTHO gene families of the PLAZA2.5 database as a gene family scaffold (Van Bel *et al*., 2012). ORTHO families provide the appropriate gene family scaffold for our analysis owing to their relatively small size. We assigned protein translations of all *Sphagnum* transcripts to ORTHO gene families using their best blastp hit (Altschul *et al.*, 1997) against all proteins in the PLAZA2.5 database with an alignment length threshold of ≥150 amino acid and a

For Peer Review

minimum similarity of 30% (Rost 1999). We kept only gene families that had at least one sequence for each of the following species: *Arabidopsis thaliana*, *Populus trichocarpa*, *Oryza indica*, *Zea mays*, *Selaginella moellendorffii*, *Physcomitrella patens*, *Volvox carterii* and *Chlamydomonas reinhardii*. We further filtered these families and kept only those in which at least two *Sphagnum* peatmoss or two non-*Sphagnum* peatmoss sequences were found because we were aiming to test whether the large-scale duplication is shared by these two groups of peatmosses. Finally, gene families with transcripts that were not found in the full K_s distributions for all species analyzed were discarded.

We generated protein guided nucleotide alignments for each gene family using TranslatorX (Abascal *et al.*, 2010) and muscle (Edgar, 2004) and filtered nucleotide alignment positions with the "-automated1" option of TrimAl (Capella-Gutierrez *et al.*, 2009). We removed sequences from the TrimAl processed alignments which consisted of more than 40% gaps. We then generated maximum likelihood gene trees for each family using RaxMLhpc (Stamatakis, 2014) with the GTRGAMM model of evolution and conducting 200 fast-bootstrap replicates. We generated gene trees for 578 gene families.

We first conducted relative dating of the duplication events by comparing the gene trees with the species tree (Fig. 1). Each gene tree was rooted using *V. carteri* and/or *C. reinhardtii* and the most recent common ancestor (MRCA) node of a given paralog pair identified in the K_s analysis was retrieved in the gene tree with its bootstrap support. A paralog pair was counted to support a specific placement of the duplication event on the species tree if all species of the paralog clade (defined by the MRCA node of the paralog pair on the gene tree) were located above the specific species tree node. We also required that the sister lineage of the specific species tree node occurs in the sister clade of the paralog clade but not in the paralog clade itself in the gene tree. We counted each unique duplication node only once per species tree; that is, multiple paralog pairs that mapped to the same duplication node on the species tree were counted only once. We discarded paralog pairs for which the paralog clade and its sister clade shared species because in such situation the duplication node could not be unambiguously defined. We conducted this analysis separately at two bootstrap thresholds (50% and 80%).

For Peer Review

To estimate the relative age of the K_s duplication peaks we enumerated the age of their duplication nodes inferred by the gene tree analysis. Although our K_s-based inference suggested the presence of four duplication peaks their separation turned out to be challenging. Therefore, we joined paralog pairs of the first two peaks (Table 2) and generated a second group from the paralog pairs of the third and fourth K_s peaks (Table 2) using the following K_s thresholds: $0.1 \leq K_s \leq 1$; $1 < K_s \leq 4$. We then enumerated the relative age estimate of the paralog pairs of these two groups for both bootstrap thresholds (50% and 80%) and for each species separately.

For each gene family we used the best maximum-likelihood gene tree to obtain an absolute age estimate for the divergence of paralog pairs with an inferred duplication node mapping to the MRCA of all peatmosses. We applied a semi-parametric penalized likelihood approach implemented in the software r8s (Sanderson 2003). The smoothing parameter was determined by cross-validation using the following constraints: MRCA of mosses minimum age: 400 my and maximum age: 450 my; the split of *Selaginella* from the rest of seed plants: 400 my; the split of *P. trichocarpa* and *A. thaliana*: 100 my. We also assigned a minimum age of 125 million years to the split of monocots and dicots (Jiao *et al.*, 2011; Vanneste *et al.*, 2014). We required trees to pass the cross-validation procedure. Finally, we determined the most likely number of duplication peaks using the EMMIX code and the Bayesian information criterion (BIC) (McLachlan *et al.*, 1999).

Functional annotation of the virtual transcripts

In order to look for potential functional bias among the duplicated genes that were retained following the putative genome duplication in Sphagnopsida, functional annotation for each putative gene was obtained using homology searches followed by Blast2GO (Conesa *et al.*, 2005) analysis. All transcripts that passed the various filters (see **Transcriptome assembly and ORF predictions)** were blasted against the NCBI non-redundant sequence database using BLASTP with the following initial cutoff values:

For Peer Review

e-value threshold of 10^{-3} and a high-scoring segment pair (HSP) length threshold of 33 amino acids. A maximum of 20 hits were kept per query sequence. GO annotations for the transcripts with hits were retrieved and assigned to each transcript using BLAST2GO and filtered using the following cutoffs: e-value filter 10^{-6}, annotation cutoff 55, GO weight 5 and Hsp-Hit Coverage cutoff 0. In a similar way, enzyme codes were also mapped to transcripts. To reduce complexity, GO terms were mapped to GO slim categories (The Gene Ontology Consortium, 2000), which we used in the final analyses.

GO enrichment analysis

In order to investigate functionality bias, GO enrichment analyses were conducted on the genes participating in the most recent large-scale duplication event shared by all species (referred to as Peak1 in Table 2). The GO bias analysis was conducted using FUNC (Prüfer *et al.*, 2007). Specifically, we tested whether transcripts inferred as being part of the large-scale duplication event are enriched for particular GO categories compared to the whole transcriptome (the background set) of each species. Analyses were run for each species separately. Because GO terms are not independent from one another we retained only the most specific significant term when redundancy was detected. We did this by running our analyses first on all GO terms and then conducting a "refinement" analysis to keep the most specific terms using the algorithm provided in the FUNC tool (Prüfer *et al.*, 2007). We accepted GO terms with a false-discovery rate smaller than 0.05 as significantly over- or under-represented after conducting 10,000 randomizations (Prüfer *et al.*, 2007).

A similar enrichment test was performed for tree-based analyses. In particular, we compared gene ontology (GO) annotations of the gene families showing signs of duplication versus the background set of all 578 gene families. GO enrichment tests were conducted for each peatmoss species separately. To functionally annotate gene families we attached the GO slim terms of genes and those of peatmoss transcripts to each family. GO slim annotation of protein coding genes was downloaded from the PLAZA2.5 data base (Van Bel *et al.*, 2012).

For Peer Review

Results

K_s-based analysis of duplication

K_s frequency plots include 1,828 gene duplicates for "F. novo-caledoniae", 2,009 for *E. inretortum*, 2,240 for *F. sericeum*, 8,571 for *S. cribrosum*, 2,625 for *S. fallax*, 4,605 for *S. lescurii*, 4,352 for *S. palustre*, 6,634 for *S. subsecundum* and 4,277 for *S. recurvum* (Figs. 2, 3). For each K_s plot, the optimal number of normal components, their means and standard deviation values estimated by the EMMIX analysis are shown in Table 2. BIC values are given in Supporting Information Table S1. For all species investigated, K_s frequency plots exhibit four distinguishable peaks at K_S ~0.26 – 0.5, K_s ~ 0.6 – 0.8, K_s ~ 1.4 – 1.8 and K_s ~ 3.2 – 4 (Figs. 2, 3; Table 2 and Supporting Information Table S1), providing evidence for at least four large-scale or genome-wide duplication events. *Eosphagnum inretortum, S. subsecundum* and S. *cribrosum* K_s plots also show a population of putative paralog pairs with a normal distribution centered around K_s of ~0.025 (Fig. 2). These components may represent pairs of very recently duplicated genes or sequencing errors that resulted in assembly of distinct transcripts with high sequence identity.

As exemplified by "F. novo-caledoniae", the shape of the K_s distribution derived from the set of paralogous gene pairs is similar, irrespective of the assembler used (Fig. 3). This confirms that any putative signal of WGD is robust to the algorithm used for assembling Illumina reads into transcripts.

In all three species-pair comparisons, the orthologs' K_s values are centered around ~0.2 while paralogs K_s values are centered around ~0.4 for the most recent duplication event (Fig. 4 A-C). Divergence of these genera is thus more recent than the duplication events themselves, which would suggest that the most recent WGD event occurred in an ancestral lineage leading to Sphagnopsida.

For Peer Review

Tree-based analysis of duplication

In total we generated 578 gene trees of which 455 and 412 supported at least one duplication event with bootstrap support ≥50% or ≥80%, respectively. Remarkably, we observed only seven gene families in which all nine investigated species had at least one paralog pair with a duplication node showing a bootstrap support ≥50%. Assuming that the large-scale duplication affected the whole genome this observation suggests highly variable retention of paralog pairs among species.

On average, 71-75% of the duplication events mapped to the MRCA of all peatmosses for both bootstrap thresholds used (Table 3). Therefore, our phylogenomic analysis confirms that all peatmosses share two large-scale duplication events. We also found that gene trees containing paralog pairs of the two most recent K_s peaks (peaks 1 and 2 in Table 2) represent duplication events predominantly (more than 80% of them) mapping to the MRCA of all peatmosses. Only a few appear to be within-species paralogs or are representatives of older duplications (Table 3, 4).

4-7% of the gene trees duplication events were inferred at the MRCA of mosses (MRCA of *P. patens* and all peatmosses) or at the MRCA of Viridiplantae (Table 3 and Fig. 5). These duplicates were predominantly located in the third and fourth peaks of the K_s distribution (Table 2) which, however, also contained a considerable number of paralogs (about 50%) with duplication events mapping to the MRCA of all peatmosses (Table 4). Altogether, both the K_s- and gene tree-based analyses suggest that at least two large-scale duplication events coincide with or predate the MRCA of peatmosses but postdate the MRCA of mosses, and these account for about 71-75% of the gene duplications investigated. In contrast, a negligible number of paralog pairs result from a duplication in the MRCA of all mosses and/or Viridiplantae (Fig. 5).

We found that divergence time of the paralog pairs with duplication event mapping to the MRCA of peatmosses could be reliably estimated for only 72 gene trees because the rest

For Peer Review

of the trees did not pass the cross validation procedure and the smoothing parameter could not be obtained. Running the EMMIX algorithm on the estimated age distribution favored one normal distribution as the most likely solution over two, with a moderate support ($BIC_{single\ peak}$= 488, age=197 my [95%CI: ±24]; $BIC_{two\ peaks}$= 494, age_1=218 my [95%CI: ±29], age_2=112 my [95%CI: ±10]).

Functional enrichment of duplicated gene sets

We used the nine GO annotated transcriptomes to ask whether similar functional gene groups have been preferentially retained after large-scale duplication. For the K_s-based analysis we only investigated the most recent duplication event shared by all species (peak 1 in Table 2). We also conducted a similar analysis by testing for functional enrichment in the gene trees with duplication events that mapped to the MRCA of all peatmosses relative to the full set of gene trees. Gene pairs preferentially retained after large-scale duplication were enriched for multiple GO slim categories (Table 5). In particular, in the K_s-based analysis we found 22, 18, and 22 GO slim categories that were overrepresented in the duplicated gene set in at least one species in the Biological process, Molecular function, and Cellular component ontologies, respectively. These numbers were similar or slightly higher in the gene-tree based analysis with 42, 13 and 21 GO slim categories. Enriched GO terms were variable among the nine species, showing an overall greater consistency across species in the gene-tree than in the K_s-based analysis (Table 5 and 6). For instance, no enriched GO slim category was shared across all nine species in the K_s-based analysis while we found six such GO slim categories in the gene-tree based analysis. Furthermore, only one GO slim term was shared by five out of the nine species in the K_s-based analysis while multiple enriched GO slim terms were shared by five or more species in the gene-tree based analysis (Table 6).

Although GO slim term sharing among species was limited, some functional categories were shared by multiple species (at least three) both in the K_s- and gene-tree based analyses (Table 5 and 6). Paralog pairs retained were frequently enriched for GO slim terms of the biological process ontology such as signal transduction, protein transport,

For Peer Review

response to abiotic and endogenous stimuli in both the K_s and gene-tree based analyses. The gene-tree based analysis also suggested the enrichment of functional categories such as transport, response to stress and to other stimuli, biosynthetic processes, cellular component organization and homeostasis. Based on the K_s analysis, retained paralogs were frequently enriched for protein kinase function of the Molecular function ontology, but many more frequently enriched categories were revealed in the gene tree-based analysis, including transporter activity, protein binding, transferase activity, catalytic activity and structural molecule activity (Table 5 and 6). Both the K_s- and gene tree-based analyses suggested that retained paralog pairs were preferentially located in protein complexes of the vacuole, nucleus, plasma membrane in the ribosome or in other organellar membranes (Table 5 and 6). These results suggest that at a certain level preferential paralog retention is driven by functional properties of genes; nevertheless, there is considerable variability among species.

Our gene-tree based analysis suggested that only seven gene families show shared duplication across all nine species investigated. Although some of these families had unknown functions others contained genes that are putative ATPase driven ABC-2 transporter genes, protein-L-isoaspartate methyltransferase genes or putative chloroplast chaperonins.

Discussion

--Does the inferred duplication coincide with the recent diversification of this group?

Studies using molecular dating suggest that although the Sphagnopsida diverged from other mosses hundreds of millions of years ago, extant *Sphagnum* diversified rather recently (ca. 20 mya; Shaw *et al.*, 2010a). Our analyses indicate that two large-scale duplication events predated the split of *Sphagnum* and non-*Sphagnum* peatmosses but postdated that of the MRCA of mosses (Fig. 3, 5 and Table 3 and 4). Therefore, the large-scale duplications must have occurred in the ancestor of all peat mosses (both non-

For Peer Review

519 *Sphagnum* and *Sphagnum*) contradicting our hypothesis that only the more speciose
520 *Sphagnum* peatmosses have experienced a large-scale duplication event. That is, our data
521 does not indicate a direct link between the large-scale duplication event and *Sphagnum*
522 diversification. Nevertheless, this does not exclude the possibility that the observed large-
523 scale duplication could have contributed to the rapid diversification of *Sphagnum*
524 peatmosses. It was recently shown that there is statistical association between
525 polyploidization and diversification in angiosperms but there is a certain lag time
526 between the polyploidzation and the diversification events (Soltis *et al.*, 2009; Schranz *et*
527 *al.*, 2012; Tank *et al.*, 2015). Therefore, it can be speculated that our results may be in
528 line with the lag hypothesis. *Sphagnum* subgenera comprise monophyletic groups that are
529 to a significant extent ecologically divergent in microhabitat within peatlands (Johnson *et*
530 *al.*, 2015). Divergent retention of duplicates among species occurring in contrasting
531 microhabitats and increased substitution rates of duplicates may have contributed to the
532 rapid radiation of *Sphagnum* peatmosses into ecologically diverse microhabitats. With
533 further data and statistical analysis this hypothesis may become testable in the near
534 future.
535
536 We also found that about 29-25% of the duplication events appear to predate the MRCA
537 of all peatmosses and map either to the MRCA of mosses or to the MRCA of
538 viridiplantae. Nevertheless, the absolute numbers of paralog pairs and gene trees
539 supporting these deep duplication events were low (Table 3 and 4) and overall
540 significance of these duplicates is unclear. If these deep duplication events turn out to be
541 genome-wide they will represent the deepest large-scale duplication event yet reported
542 for land plants. At present, the deepest large-scale duplication event recovered is shared
543 by all seed plants but not by lycophytes (Jiao *et al.*, 2011).
544
545 *--The WGD shared by all peatmosses and the Cretaceous-Paleogene mass extinction*
546 It has been proposed that large-scale duplications in multiple land plant lineages are
547 chronologically clustered around the time of the end-Cretaceous mass extinction (Fawcett
548 *et al.*, 2009; Vanneste *et al.*, 2014). Lineages with recent genome duplications might have
549 coped better with the severe environmental conditions, or duplications may have

For Peer Review

facilitated an immediate and rapid response to severe environmental stress (Visscher *et al.*, 2004). The question arises if genome duplication in the Sphagnopsida also occurred at this time. The whole-genome duplication revealed in the *P. patens* genome is assumed to have coincided with the end-Cretaceous mass extinction and was inferred to occurred 60-70 million years ago (Fawcett *et al.*, 2009; Vanneste *et al.*, 2014). Our absolute dating analysis suggests that the duplication found in all peatmoss species is considerably older than that (197 ± 24 my). Because we constrained nodes with the same ages as in studies dating the WGD in *P. patens* (Fawcett *et al.*, 2009; Vanneste *et al.*, 2014) we believe that our estimate is reliable. Therefore, our inference suggests that the large-scale duplication is chronologically incongruent with the end-Cretaceous mass extinction event and it might have been driven by other ecological/climatological/physiological factors. Nevertheless, other events yet to be discovered in bryophytes may turn out to coincide chronologically with large-scale genome duplications clustered around the Cretaceous-Paleogene mass extinction (Vanneste *et al.*, 2014).

--*Gene retention is variable across species in peatmosses*

Most gene duplicates after large-scale duplications are typically lost and only a small proportion is retained over longer periods of time (Vision *et al.*, 2000; Blanc & Wolfe, 2004; Paterson *et al.*, 2004; Van de Peer & Meyer, 2005; Cui *et al.*, 2006; Jaillon *et al.*, 2007; Shi *et al.*, 2010; Barker *et al.*, 2012; Jiao *et al.*, 2012; Amborella Genome Consortium, 2013). Multiple, non-exclusive processes may explain the preferential retention of gene duplicates, including gene dosage requirements (Conant *et al.*, 2014). The relative gene dosage hypothesis suggests that genes that are part of multiprotein complexes are preferentially retained because proper function can be only achieved when components are produced in the right stoichiometry (Papp *et al.*, 2003; Freeling & Thomas, 2006; Edger & Pires, 2009; Freeling, 2009; Birchler & Veitia, 2011; Conant *et al.*, 2014). Previous findings seem to conform to this pattern (Vision *et al.*, 2000; Blanc & Wolfe, 2004; Paterson *et al.*, 2004; Van de Peer & Meyer, 2005; Cui *et al.*, 2006; Jaillon *et al.*, 2007; Barker *et al.*, 2008; Shi *et al.*, 2010 ; Jiao *et al.*, 2011 ; Amborella Genome

For Peer Review

581 Consortium, 2013). If this applies to the peatmoss species investigated here we would
582 expect that similar categories would be preferentially retained in all or most of the
583 species. Our data are generally in favor of the relative gene dosage hypothesis. Both the
584 K_s- and the gene tree-based analyses show that the preferentially retained gene set is
585 enriched for proteins that are part of the ribosome or of ion channels and signal
586 transduction pathways. This is in line with previous studies showing that genes that are
587 highly connected and/or are part of macromolecular complexes or stoichiometric
588 pathways are usually preferentially retained after gene duplication in angiosperms (Hakes
589 *et al.*, 2007; Freeling, 2009; Bekaert *et al.*, 2011; Rodgers-Melnick *et al.*, 2012).
590
591 We also found patterns of preferential retention that cannot be explained by the relative
592 dosage hypothesis. These include response to endogenous stimuli or stress, defense
593 response, transporter activity, and cellular homeostasis. We argue that preferential
594 retention of these functional categories of genes may be explained by the general biology
595 of peatmosses. For instance, peatmosses are the dominant components of peatlands,
596 which are nutrient-poor environments with an acidic pH, unusual metal ion
597 concentrations, and metal accumulation (Saxena, 2006; Saxena & Saxena, 2012). It is
598 thus possible that preferential retention of genes with transporter function or importance
599 in stress response has contributed to the success of peatmosses in their special
600 environment.
601
602 Importantly, we found relatively few functional categories that were preferentially
603 represented among retained genes across species compared to another study using a
604 similar experimental design (Barker *et al.*, 2008). This finding can be explained at least in
605 two ways. First, paralog retention may be related to the ecological niches of the different
606 species and thus the absence of a common pattern across species might be explained by
607 local adaptation rather than by relative gene dosage. The latter hypothesis is consistent
608 with the observation that *Sphagnum* species sort predictably along niche microhabitat
609 gradients within peatlands and niche differentiation likely played an important role in
610 peatmoss diversification (Rydin & Juglum, 2006; Johnson *et al.*, 2015a, 2015b).

For Peer Review

Nevertheless, this hypothesis can be only assessed by careful experimental work on the functional divergence of groups of genes across species.

Alternatively, it could be argued that dissimilar enrichment of GO terms among the species in our study may be an artifact because of differential expression of genes or paralogs across species. Although this bias could certainly affect our results, the study design was intended to minimize this effect. We extracted RNA using the actively growing part of the plant (capitulum), which is similarly structured in all species investigated. We also repeated our analyses using the annotated gene set of the *P. patens* v1.6 genome as the background set (results not shown), but this did not change the pattern we observed in the diversity of GO categories for retained paralogues. Although environmental effects on gene expression patterns cannot be discounted, other studies using a similar design, sampling only leaves from multiple species, have found that in contrast to our results similar gene categories were preferentially retained as duplicates across species (Barker *et al.*, 2008; Barker *et al.*, 2012). We therefore suggest that expression differences among species are not a likely explanation for the species-specific patterns we observed.

Acknowledgments

This research was supported by NSF grant no. DEB-0918998 to A.J. Shaw and Blanka Shaw and by an SNSF Ambizione (#131726), an SNSF project grant, and a grant from the URPP in Systems Biology/Functional Genomics and Evolution in Action to P Szövényi. Sequencing and analysis for *S. fallax* was supported by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U. S. Department of Energy. The 1000 Plants Project (1KP) initiative is funded by the Alberta Ministry of Innovation and Advanced Education, Alberta Innovates Technology Futures' Innovates Centers of Research Excellence program, Musea Ventures, and BGI-Shenzhen. We thank Michael McKain and Jim Leebens-Mack for their help in data analysis. Comments of two anonymous reviewers on an earlier version of this manuscript are also greatly acknowledged.

For Peer Review

642 **Author contribution**

643

644 P.SZ., N.D. and J.A.S planned and designed the research. P.SZ., N.D., M.J. and C.R
645 performed experiments. P.SZ., N.D. and M.J. analysed data. P.SZ.,N.D.,J.A.S, M.J.,
646 C.R. and D.W. wrote the manuscript.

647

648

649

650

651 **References**

652

653 **Abascal F, Zardoya R, Telford MJ. 2010.** TranslatorX: multiple alignment of
654 nucleotide sequences guided by amino acid translations. *Nucleic Acids Research* **38**:W7-
655 13.
656 **Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.**
657 **1997.** Gapped BLAST and PSI-BLAST: a new generation of protein database search
658 programs. *Nucleic Acids Research* **25**:3389-3402.
659 **Amborella Genome Consortium. Chamala S, Chanderbali A, Der J, Lan T, Walts B,**
660 **Albert V, dePamphilis C, Leebens-Mack J, Rounsley S, Schuster S et al. 2013.** The
661 complete nuclear genome of *Amborella trichopoda*: an evolutionary reference genome
662 for the angiosperms. *Science* **342**: art. no. 6165.
663 **Arrigo N, Barker MS. 2015.** Rarely successful polyploids and their legacy in plant
664 genomes. *Current Opinion in Plant Biology* **15**:140-146. doi: 10.1016/j.pbi.2012.03.010.
665 **Barker MS, Baute GJ, Liu SL. 2012.** Duplications and turnover in plant genomes. In:,
666 J. F. Wendel, ed. *Plant Genome Diversity.* Vienna, Austria: Springer, Volume 1,155-169.
667 **Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore W, Knapp SJ, Rieseberg**
668 **LH. 2008.** Multiple paleopolyploidizations during the evolution of the Compositae reveal
669 parallel patterns of duplicate gene retention after millions of years. *Molecular Biology*
670 *and Evolution* **25**:2445–2455.

For Peer Review

671 **Barker MS, Vogel H, Schranz ME. 2009.** Paleopolyploidy in the Brassicales: analyses
672 of the cleome transcriptome elucidate the history of genome duplications in Arabidopsis
673 and other Brassicales. *Genome Biology and Evolution* **1**:391–399.
674 **Bekaert M, Edger PP, Pires JC, Conant GC. 2011.** Two phase resolution of polyploidy
675 in the *Arabidopsis* metabolic network gives rise to relative and absolute dosage
676 constraints. *Plant Cell* **23**:1719–1728.
677 **Birchler JA, Veitia RA. 2011.** Protein-protein and protein-DNA dosage balance and
678 differential paralog transcription factor retention in polyploids. *Frontiers in Plant*
679 *Sciences* **2**:64. doi: 10.3389/fpls.2011.00064.
680 **Blanc G, Wolfe KH. 2004.** Widespread paleopolyploidy in model plant species inferred
681 from age distributions of duplicate genes. *Plant Cell* **16**:1667–1678.
682 **Bowers JE, Chapman BA, Rong JK, Paterson AH. 2003.** Unraveling angiosperm
683 genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*
684 **422**:433–438.
685 **Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang X, Mudge J,**
686 **Vasdewani J, Scheix T et al. 2006.** Legume genome evolution viewed through the
687 *Medicago truncatula* and *Lotus japonicus* genomes. *Proceedings of the National*
688 *Academy of Sciences, USA* **103**:14959–14964.
689 **Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009.** TrimAl: a tool for
690 automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*
691 **25**:1972-1973.
692 **Carretero-Paulet L, Fares MA. 2012.** Evolutionary dynamics and functional
693 specialization of plant paralogs formed by whole and small-scale genome duplications.
694 *Molecular Biology and Evolution* **29**:3541-3551.
695 **Chang Y, Graham SW. 2011.** Inferring the higher-order phylogeny of mosses
696 (Bryophyta) and relatives using a large, multigene plastid data set. *American Journal of*
697 *Botany* **98**: 839-849.
698 **Chen EC, Buen Abad Najar C, Zheng C, Brandts A, Lyons E, Tang H, Carretero-**
699 **Paulet L, Albert VA, Sankoff D. 2013.** The dynamics of functional classes of plant
700 genes in rediploidized ancient polyploids. *BMC Bioinformatics* **14 Suppl 15**:S19. doi:
701 10.1186/1471-2105-14-S15-S19.

For Peer Review

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology* **19**:91-98. doi: 10.1016/j.pbi.2014.05.008.

Conant GC. 2014. Comparative genomics as a time machine: how relative gene dosage and metabolic requirements shaped the time-dependent resolution of yeast polyploidy. *Molecular Biology and Evolution* **31**:3184-3193.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Montserrat R. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674-3676.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* **16**:738–749.

Doyle JJ. 2012. Polyploidy in Legumes. In: Soltis PS, Soltis DE, eds. *Polyploidy and genome evolution*. New York, USA: Springer, 147–180.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**:1792-1797.

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research* **17**:699-717.

Escudero M, Martín-Bravo S, Mayrose I, Fernández-Mazuecos M, Fiz-Palacios O, Hipp AL, Pimentel M, Jiménez-Mejías P, Valcárcel V, Vargas P et al. 2014 Karyotypic changes through dysploidy persist longer over evolutionary time than polyploid changes. *Plos One* **9**:e85266. doi: 10.1371/journal.pone.0085266.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proceedings of the National Academy of Science*s **106**:5737-5742. doi: 10.1073/pnas.0900906106.

Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. 2008. Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. *Genome Research* **18**:1924-1937.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review in Plant Biology* **60**:433-453.

For Peer Review

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research* **16**:805-814.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**:3150-3152. doi: 10.1093/bioinformatics/bts565.

Gao B, Zhang D, Li X, Yang H, Wood AJ. 2014. De novo assembly and characterization of the transcriptome in the desiccation-tolerant moss *Syntrichia caninervis*. *BMC Research Notes* **7**:490. doi: 10.1186/1756-0500-7-490.

Goffinet B, Shaw AJ. 2009. *Bryophyte Biology*, Cambridge University Press

Gorham E. 1991. Northern Peatlands: Role in the Carbon Cycle and Probable Responses to Climatic Warming. *Ecological Applications* **1**:182–195.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. 2001. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology* **29**:644-652.

Greb SF, DiMichele WA, Gastaldo RA. 2006. Evolution and importance of wetlands in earth history. In: Greb SF, DiMichele WA eds. *Wetlands through time.* Geological Society of America*,* Special Paper 399, 1-40,

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**:1494-1512.

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biology* **8**: R209.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**:463–467.

Jiang WK, Liu YL, Xia EH, Gao LZ. 2013. Prevalent role of gene features in determining evolutionary fates of whole-genome duplication duplicated genes in flowering plants. *Plant Physiology* **161**:1844-1861.

For Peer Review

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biology* **13**:R3.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**:97–100.

Johnson MG, Shaw AJ. 2015. Genetic diversity, sexual condition, and microhabitat preference determine mating patterns in *Sphagnum* (Sphagnaceae) peatmosses. *Biological Journal of the Linnean Society*: doi: 10.1111/bij.12497.

Johnson, MG, Granath G, Teemu T, Pouliot R, Stenøien HK, Rochefort L, Rydin H, Shaw AJ. 2015. Evolution of niche preference in *Sphagnum* peat mosses. *Evolution* **69**: 90-103.

Karlin EF, Giusti MM, Lake RA, Boles SB, Shaw AJ. 2010. Microsatellite analysis of *Sphagnum centrale*, *S. henryense*, and *S. palustre* (Sphagnaceae). *The Bryologist* **113**:90-98.

Liu SL, Baute GJ, Adams KL. 2011. Organ and cell type-specific complementary expression patterns and regulatory neofunctionalization between duplicated genes in *Arabidopsis thaliana*. *Genome Biology and Evolution* **3**:1419-1436. doi: 10.1093/gbe/evr114.

Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D et al. 2008. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiology* **148**:1772–1781.

Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences, USA* **107**:9270-9274.

Matasci N, Hung LH, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M et al. 2014. Data access for the 1,000 Plants (1KP) project. *GigaScience* **3**:17.

Mayrose I, Zhan SH, Rothfels CJ, Arrigo N, Barker MS, Rieseberg LH, Otto SP. 2014. Methods for studying polyploid diversification and the dead end hypothesis: a

For Peer Review

reply to Soltis et al. (2014). *New Phytologist* **206**:27-35.

Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, Otto SP. 2011. Recently formed polyploid plants diversify at lower rates. *Science* **333**:1257.

McKain MR, Wickett N, Zhang Y, Ayyampalayam S, McCombie WR, Chase MW, Pires JC, dePamphilis CW, Leebens-Mack J. 2012. Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). *American Journal of Botany* **99**:397–406.

McLachlan G, Peel D, Basford KE, Adams P. 1999. The EMMIX algorithm for the fitting of normal and t-components. *Journal of Statistical Software* **4**:i02.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KLT et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* L.). *Nature* **452**: 991–996.

Ohno S. 1970. Evolution by Gene Duplication, Springer-Verlag, Heidelberg.

Ohta T. 1987. Simulating evolution by gene duplication. *Genetics* **115**:207–213.

Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**:194-197.

Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences, USA* **101**:9903–9908.

Ponting CP, Russell RR. 2002. The natural history of protein domains. *Annual Review of Biophysics and Biomolecular Structure* **31**:45–71.

Prüfer K, Muetzel B, Do HH, Weiss G, Khaitovich P, Rahm E, Pääbo S, Lachmann M, Enard W. 2007. FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* **8**:41.

Rensing SA, Ick J, Fawcett JA, Lang D, Zimmer A, Van de Peer Y, Reski R. 2007. An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evolutionary Biology* **7**:130. doi:10.1186/1471-2148-7-130.

For Peer Review

Rensing SA. 2014. Gene duplication as a driver of plant morphogenetic evolution. *Current Opinion in Plant Biology* **17**:43-48.

Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, Difazio SP. 2012. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. *Genome Research* **22**:95–105.

Rost B. 1999. Twilight zone of protein sequence alignments. *Protein Engineering* **12**:85-94.

Rydin H, Jeglum J. 2006. *The Biology of Peatlands*. New York, NY, USA: Oxford University Press.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**:301–302.

Saxena A, Saxena A. 2012. Bioaccumulation and glutathione-mediated detoxification of copper and cadmium in *Sphagnum squarrosum* Crome Samml. *Environmental Monitoring and Assessment* **184**:4097-4103. doi: 10.1007/s10661-011-2246-9.

Saxena A. 2006. Seasonal pattern of metal bioaccumulation and their toxicity on *Sphagnum squarrosum*. *Journal of Environmental Biology* **27**:71-75.

Scarpino SV, Levin DA, Meyers LA. 2014. Polyploid formation shapes flowering plant diversity. *The American Naturalist* **184**:456-465.

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47**:868-876.

Schnable JC, Pedersen BS, Subramaniam S, Freeling M. 2012. Dose-sensitivity, conserved non-coding sequences, and duplicate gene retention through multiple tetraploidies in the grasses. *Frontiers in Plant Sciences* **2**:2.

Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: The WGD radiation lag-time model. *Current Opinion in Plant Biology* **15**:147–153.

Shaw AJ, Cox CJ, Buck WR, Devos N, Buchanan AM, Cave L, Seppelt R, Shaw B, Larraín J, Andrus R et al. 2010b. Newly resolved relationships in an early land plant lineage: Bryophyta class Sphagnopsida (peat mosses). *American Journal of Botany* **97**: 1511-1531.

For Peer Review

Shaw AJ, Devos N, Cox CJ, Boles SB, Shaw B, Buchanan AM, Cave L, Seppelt R. 2010a. Peatmoss (*Sphagnum*) diversification associated with Miocene Northern Hemisphere climatic cooling? *Molecular Phylogenetics and Evolution* **55**:1139–1145.

Shi T, Huang H, Barker MS. 2010. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. *Annals of Botany* **106**:497–504.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* **96**:336–348.

Soltis DE, Visger CJ, Soltis PS. 2014. The polyploidy revolution then…and now: Stebbins revisited. *American Journal of Botany* **101**:1057-1078.

Stamatakis A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* **30**:1312-1313.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* **34**(Web Server issue):W609-W612. doi:10.1093/nar/gkl315.

Szövényi P, Perroud PF, Symeonidi A, Stevenson S, Quatrano R, Rensing S, Cuming A, McDaniel SF. 2015. De novo assembly and comparative analysis of the *Ceratodon purpureus* transcriptome. *Molecular Ecology Resources* **15**:203-215.

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, Brown JW, Sessa EB, Harmon LJ 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist* **207**:454-467.

Tayale A, Parisod C. 2013. Natural pathways to polyploidy in plants and consequences for genome reorganization. *Cytogenetic and Genome Research* **140**:79–96.

The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* **25**:25-29.

The UniProt Consortium. 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research* **42**:D191-D198.

Turetsky MR, Crow SE, Evans RJ, Vitt DH, Wieder RK. 2008. Trade-offs in resource allocation among moss species control decomposition in boreal peatlands. *Journal of Ecology* **96**:1297-1305.

For Peer Review

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**:1596–1604.

Van Bel M, Proost S, Wischnitzki A, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K. 2012. Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiology* **158**:590-600.

Van de Peer Y, Meyer A. 2005. Large-scale gene and ancient genome duplications. In: Gregory TR, ed. *The evolution of the genome*. San Diego, CA, USA: Elsevier Academic Press, 329–368.

Van de Peer Y. 2011. A mystery unveiled. *Genome Biology* **12**:113–114.

Vanneste K, Maere S, Van de Peer Y. 2014. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philosophical Transactions Royal Society London B Biological Sciences* **369**. pii: 20130353. doi: 10.1098/rstb.2013.0353.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nature Review Genetics* **10**:725-732. doi: 10.1038/nrg2600.

Vanneste K, Van de Peer Y, Maere S. 2013. Inference of genome duplications from age distributions revisited. *Molecular Biology and Evolution* **30**:177-190.

Vasander H, Kettunen A. 2006. Carbon in boreal peatlands. In: Wieder RK, Vitt D.H. eds. *Boreal Peatland Ecosystems*. Heidelberg,Germany: Springer, 165-194.

Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* **290**:2114–2117.

Visscher H, Looy CV, Collinson ME, Brinkhuis H, van Konijnenburg-van Cittert JH, Kurschner WM, Sephton MA. 2004. Environmental mutagenesis during the end-Permian ecological crisis. *Proceedings of the National Academy of Sciences USA* **101**:12952–12956.

Walsh JB. 1995. How often do duplicated genes evolve new functions? *Genetics* **39**:421–428.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants.

For Peer Review

Proceedings of the National Academy of Sciences of the United States of America 1–21. doi:10.1073/pnas.1323926111

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S et al. 2014. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**:1660-1666.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**:32-43.

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. 2006. KaKs Calculator: Calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**:259-263

Zimmer AD, Lang D, Buchta K, Rombauts S, Nishiyama T, Hasebe M, Van de Peer Y, Rensing SA, Reski R. 2013. Reannotation and extended community resources for the genome of the non-seed plant *Physcomitrella patens* provide insights into the evolution of plant gene structures and functions. *BMC Genomics* **14**:498. doi: 10.1186/1471-2164-14-498.

Table S1 Bayesian Information Criterion (BIC) values when fitting 1 to 8 components to the Ks distributions by EMMIX.

Fig. 1. Species tree depicting the phylogenetic relationship of the species used in the gene tree-based analysis. Most recent common ancestors (MRCA) are shown in gray circles. Branch lengths are not to scale.

Fig. 2. K_s frequency plots for paralogous gene pairs from nine *Sphagnopsida* species. K_s distribution components estimated by EMMIX (see Materials and Methods) are superimposed on the histogram for each K_s plot. Components shared by the majority of the species are numbered according to Table 2.

Fig. 3. K_s frequency plots for paralogous gene pairs in "Flatbergium novo-caledoniae". K_s values are derived from the transcriptome assembled with Trinity (A) and

For Peer Review

946 SOAPdenovo-Trans (B). K_s distribution components inferred by EMMIX (see Materials
947 and Methods) are superimposed on the histogram for each K_s plot.
948
949 Fig. 4. K_s frequency plots for paralogous and orthologous gene pairs from (A) "F. novo-
950 caledoniae" and *S. lescurii*, (B) *E. inretortum* and *S. recurvum*, and (C) *F. sericeum* and *S.*
951 *palustre.* In all species pairs, the K_s distribution for paralogs exhibits a mode greater than
952 the modal K_s values for putative orthologs.

953 Fig. 5. Inferred location of duplication events discovered by the analysis of peatmoss
954 paralog pairs. Duplications are indicated by Gaussian curves filled with gray color.
955 Height of curves refers to the relative number of duplication events discovered. The
956 question mark above the two most recent duplication events indicates that it is currently
957 unclear whether there were one or two duplication events. Age of nodes is taken from
958 Shaw et al. (2010a) and from timetree (www.timetree.org).
959
960

For Peer Review

Table 1. Voucher information and number of transcripts for species used in the K_s analysis. For "F. novo-caledoniae" transcript numbers are given both for the Trinity and SOAPdenovo-Trans assemblies (separated by a forward slash)

Species	Voucher	Locality	ArrayExpress /SRA/Bioproject accession #	Total # transcripts
"F. novo-caledoniae"	B. Shaw 17606 (DUKE)	New Caledonia	E-MTAB-4112	62531/12382
E. inretortum	B. Shaw 18836 (DUKE)	Chile	E-MTAB-4112	38755
F. sericeum	Ho, B.C. & Yong, K.T. 13-116 (SING)	Peninsular Malaysia	E-MTAB-4112	56092
S. cribrosum	Shaw s.n 12/10/2011 (DUKE)	NC USA	E-MTAB-4112	81809
S. fallax	D. Weston s.n	MN USA	PRJNA307187	12758
S. lescurii	Shaw s.n 2013/1 (DUKE)	NC USA	ERX337183	17979
S. palustre	Rothfels 4143 (DUKE)	NC USA	SRA319444	24135
S. recurvum	Rothfels 4144 (DUKE)	NC USA	SRA319891	21302
S. subsecundum	Shaw s.n 07/4/2011 (DUKE)	ME USA	E-MTAB-2482	48844

For Review

Table 2. Mean and standard deviation (in brackets) of the components (normal distributions fitted) found by EMMIX. For "F. novo-caledoniae" components inferred using the SOAPdenovo-Trans and trinity assemblies are also shown (values are separated by a forward-slash). Components are manually aligned across species and sorted in an increasing order. NP indicates that no corresponding peak was found; + refers to a significant secondary peak in the distribution. Components shared by the majority of the species are labelled as Peak 1-4.

Species	**peak**	**peak**	**Peak 1**	**Peak 2**	**Peak 3**	**peak**	**peak**	**Peak4**
"F. novo-caledoniae"	NP/0.045(0.00032)	NP/0.099(0.0015)	0.43(0.01)/ 0.34(0.02)	0.71(0.034)/ 0.79(0.08)	1.69(0.298)/ 1.88(0.22)	NP	NP	3.52(0.540)/ 3.7(0.38)
E. inretortum	0.05(5e-4)	0.13(0.003)	0.49(0.05)		1.7(0.40)	NP	NP	3.7(0.40)
F. sericeum	0.044(2e-4)	0.09(0.0012)	0.263(0.012)	0.62(0.05)	1.8(0.40)	NP	NP	3.79(0.32)
S. cribrosum	0.046(0.0004)	0.10(0.0016)	0.30(0.01)	0.59(0.04)	1.82(0.35)	NP	NP	3.71(0.29)
S. fallax	NP	NP	0.30(0.033)	0.76(0.08)	1.66(0.21)	2.9(0.45) + 2.71(0.50)	3.06(0.40)	3.5(0.30)
S. lescurii	NP	0.09(0.0015)	0.42(0.03)		1.4(0.26)	2.1(0.4)	2.88(0.71)	3.74(0.14)
S. palustre	NP	0.08(0.0023)	0.33(0.016)	0.66(0.043)	1.68(0.33)	NP	3.06(0.98)	3.53(0.22)
S. recurvum	NP	0.08(0.0016)	0.36(0.018)	0.71(0.055)	1.62(0.22)	NP	2.91(0.69)	3.87(0.10)
S. subsecundum	0.05(0.0006)	0.13(0.003)	0.38(0.03)		1.58(0.38)	NP	NP	3.19(0.73) + 4.1(0.16)

For Peer Review

Table 3 Number and relative proportion of gene trees and paralog pairs supporting duplication events at different nodes of the species tree. Results are provided for two levels of stringency expressed as the bootstrap support of the duplication node (50% and 80 %).

Stringency	Duplication node	Species																			
		F. novo-caledoniae		*E. inretortum*		*F. sericeum*		*S. cribrosum*		*S. fallax*		*S. lescurii*		*S. palustre*		*S. recurvum*		*S. subsecundum*		AVERAGE	
bootstrap ≥ 50%																					
Total number of paralog pairs assessed		346		384		431		414		185		167		149		144		257			
	Within-species	2	4.26%	2	2.90%	2	3.08%	15	21.74%	9	24.32%	5	12.20%	5	12.82%	5	12.20%	7	16.28%		**12.20%**
	MRCA of peat mosses	40	85.11%	61	88.41%	52	80.00%	43	62.32%	26	70.27%	28	68.29%	26	66.67%	31	75.61%	21	48.84%		**71.72%**
	MRCA of mosses	3	6.38%	4	5.80%	6	9.23%	8	11.59%	2	5.41%	4	9.76%	3	7.69%	3	7.32%	6	13.95%		**8.57%**
	MRCA of viridiplantae	2	4.26%	2	2.90%	5	7.69%	3	4.35%	0	0.00%	4	9.76%	5	12.82%	2	4.88%	9	20.93%		**7.51%**
bootstrap ≥ 80%																					
Total number of paralog pairs assessed		257		294		323		283		121		101		96		86		186			
	Within-species	1	2.70%	1	1.79%	2	3.85%	8	15.38%	7	23.33%	2	6.25%	4	12.50%	4	12.12%	3	9.09%		**9.67%**
	MRCA of peat mosses	32	86.49%	53	94.64%	42	80.77%	38	73.08%	22	73.33%	24	75.00%	21	65.63%	26	78.79%	16	48.48%		**75.13%**
	MRCA of mosses	2	5.41%	1	1.79%	3	5.77%	3	5.77%	1	3.33%	2	6.25%	2	6.25%	1	3.03%	5	15.15%		**5.86%**
	MRCA of viridiplantae	2	5.41%	1	1.79%	5	9.62%	3	5.77%	0	0.00%	4	12.50%	5	15.63%	2	6.06%	9	27.27%		**9.34%**

For Peer Review

Table 4 Number and relative proportion of gene trees supporting duplication events at different nodes of the species tree. Results are provided for two ranges of paralog divergence (Ks range) and for two levels of stringency expressed as the bootstrap support of the duplication node.

Stringency	Ks range	Duplication node	Species																		
			F. novo-caledoniae		*E. inretortum*		*F. sericeum*		*S. cribrosum*		*S. fallax*		*S. lescurii*		*S. palustre*		*S. recurvum*		*S. subsecundum*		AVERAGE
bootstrap ≥ 50%																					
	0<Ks≤1																				
		Within-species	2	5.88%	2	4.44%	2	4.65%	15	28.85%	9	31.03%	5	20.00%	6	26.09%	3	13.04%	8	33.33%	**18.59%**
		MRCA of peat mosses	32	94.12%	43	95.56%	41	95.35%	37	71.15%	20	68.97%	20	80.00%	17	73.91%	19	82.61%	15	62.50%	**80.46%**
		MRCA of mosses	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	1	4.17%	**0.46%**
		MRCA of viridiplantae	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	1	4.35%	0	0.00%	**0.48%**
	1<Ks≤4																				
		Within-species	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	1	6.67%	0	0.00%	**0.74%**
		MRCA of peat mosses	10	66.67%	22	78.57%	13	54.17%	9	45.00%	8	80.00%	9	52.94%	9	52.94%	8	53.33%	8	34.78%	**57.60%**
		MRCA of mosses	3	20.00%	4	14.29%	6	25.00%	8	40.00%	2	20.00%	4	23.53%	3	17.65%	2	13.33%	6	26.09%	**22.21%**
		MRCA of viridiplantae	2	13.33%	2	7.14%	5	20.83%	3	15.00%	0	0.00%	4	23.53%	5	29.41%	4	26.67%	9	39.13%	**19.45%**
bootstrap ≥ 80%																					
	0<Ks≤1																				
		Within-species	1	3.57%	1	2.70%	1	2.78%	8	19.51%	7	28.00%	2	10.53%	5	27.78%	4	14.81%	5	23.81%	**14.83%**
		MRCA of peat mosses	27	96.43%	36	97.30%	35	97.22%	33	80.49%	18	72.00%	17	89.47%	13	72.22%	22	81.48%	15	71.43%	**84.23%**
		MRCA of mosses	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	1	4.76%	**0.53%**
		MRCA of viridiplantae	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	1	3.70%	0	0.00%	**0.41%**
	1<Ks≤4																				
		Within-species	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	1	6.67%	0	0.00%	**0.74%**
		MRCA of peat mosses	9	69.23%	19	82.61%	7	46.67%	7	53.85%	6	85.71%	8	57.14%	8	53.33%	8	53.33%	5	26.32%	**58.69%**
		MRCA of mosses	2	15.38%	2	8.70%	3	20.00%	3	23.08%	1	14.29%	2	14.29%	2	13.33%	2	13.33%	5	26.32%	**16.52%**
		MRCA of viridiplantae	2	15.38%	2	8.70%	5	33.33%	3	23.08%	0	0.00%	4	28.57%	5	33.33%	4	26.67%	9	47.37%	**24.05%**

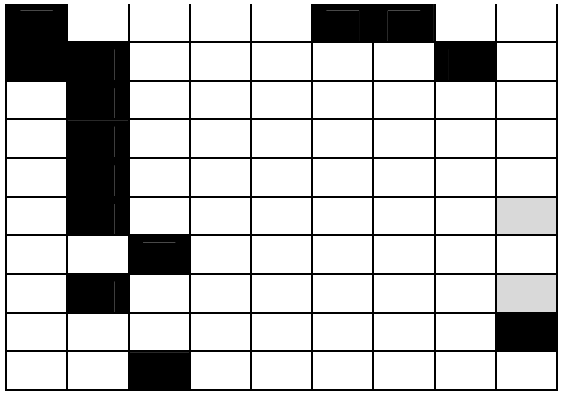
For Peer Review

Table 5. GO slim categories overrepresented in the most recent duplication event (referred to as peak 1 in table 2) shared by all species investigated. Only GO slim terms overrepresented at least in one species are shown. Significantly overrepresented terms (false-discovery rate ≤ 0.05) are shown in black and underrepresented terms in grey. Non-significant terms are in white. Description for terms overrepresented in at least three species is highlighted.

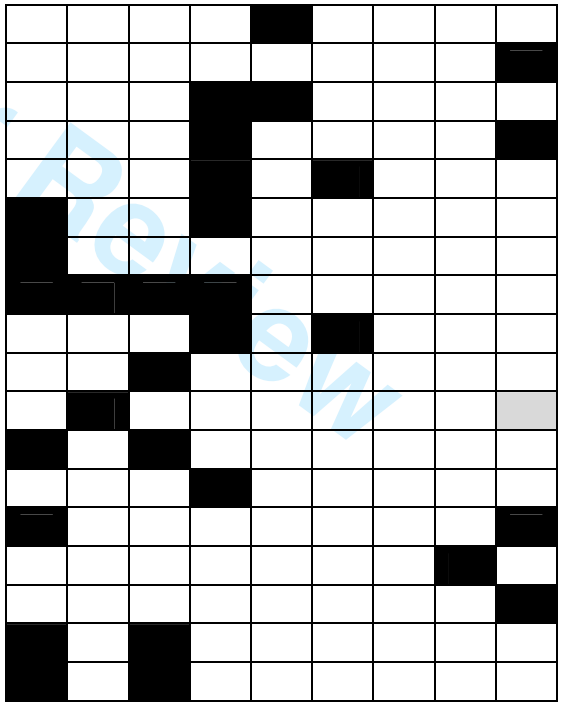
GO ID	Description	*"F. novo-caledoniae"*	*E. inretortum*	*F. sericeum*	*S. cribrosum*	*S. fallax*	*S. lescurii*	*S. palustre*	*S. recurvum*	*S. subsecundum*
Biological process ontology										
GO:0006091	generation of precursor metabolites and energy									
GO:0006259	DNA metabolic process									
GO:0006464	cellular protein modification process									
GO:0006810	transport									
GO:0006950	response to stress									
GO:0007165	**signal transduction**									
GO:0007275	multicellular organismal development									
GO:0008219	cell death									
GO:0008283	cell proliferation									
GO:0009607	response to biotic stimulus									
GO:0009628	**response to abiotic stimulus**									
GO:0009653	anatomical structure morphogenesis									

For Peer Review

GO:0009719 **response to endogenous stimulus**
GO:0015031 **protein transport**
GO:0016032 viral reproduction
GO:0016043 cellular component organization
GO:0016049 cell growth
GO:0019538 protein metabolic process
GO:0044238 primary metabolic process
GO:0044267 cellular protein metabolic process
GO:0065007 biological regulation
GO:0071704 organic substance metabolic process



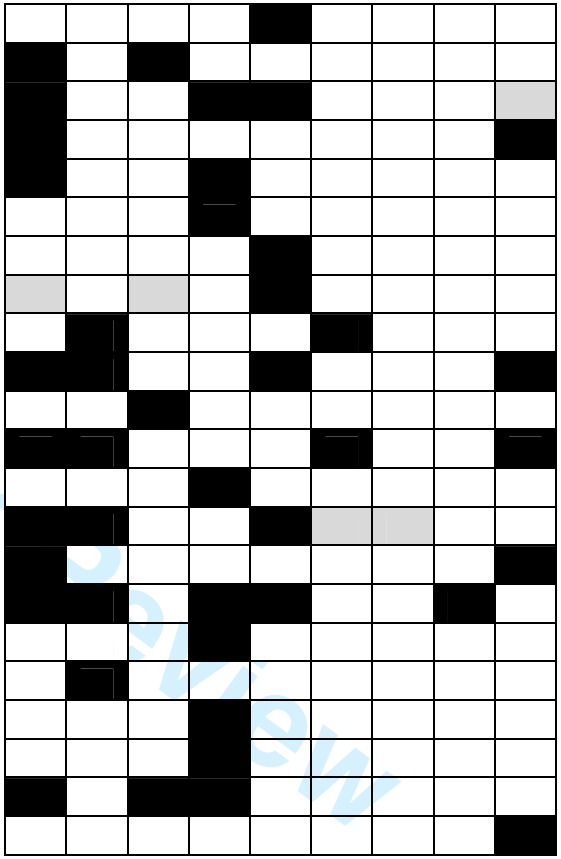
Molecular function ontology
GO:0000166 nucleotide binding
GO:0001071 nucleic acid binding transcription factor activity
GO:0003676 nucleic acid binding
GO:0003677 DNA binding
GO:0003682 chromatin binding
GO:0003700 sequence-specific DNA binding transcription factor activity
GO:0003774 motor activity
GO:0004672 **protein kinase activity**
GO:0004721 phosphoprotein phosphatase activity
GO:0004871 signal transducer activity
GO:0005198 structural molecule activity
GO:0005515 protein binding
GO:0008092 cytoskeletal protein binding
GO:0030234 enzyme regulator activity
GO:0030246 carbohydrate binding
GO:0042393 histone binding
GO:0097159 organic cyclic compound binding
GO:1901363 heterocyclic compound binding



For Peer Review

Cellular component ontology

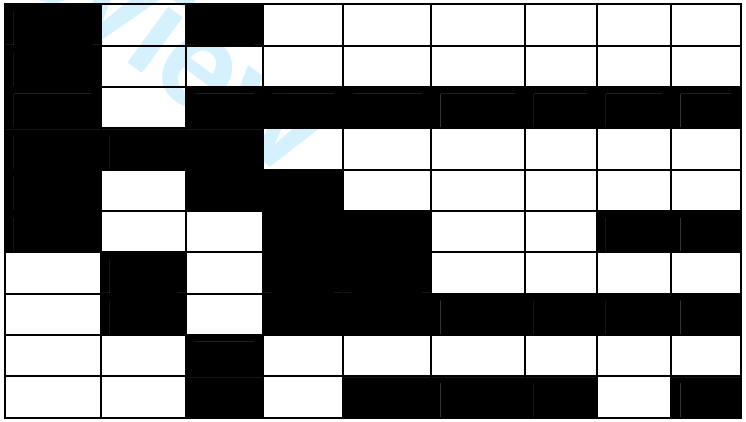
GO:0005576 extracellular region
GO:0005618 cell wall
GO:0005634 **nucleus**
GO:0005635 nuclear envelope
GO:0005654 nucleoplasm
GO:0005694 chromosome
GO:0005730 nucleolus
GO:0005739 mitochondrion
GO:0005768 endosome
GO:0005773 **vacuole**
GO:0005783 endoplasmic reticulum
GO:0005794 **Golgi apparatus**
GO:0005829 cytosol
GO:0005840 **ribosome**
GO:0005856 cytoskeleton
GO:0005886 **plasma membrane**
GO:0009579 thylakoid
GO:0016023 cytoplasmic membrane-bounded vesicle
GO:0031981 nuclear lumen
GO:0043232 intracellular non-membrane-bounded organelle
GO:0043234 **protein complex**
GO:0071944 cell periphery



For Peer Review

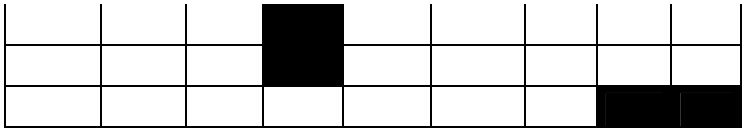
Table 6 GO slim categories overrepresented in the gene trees with one or more duplication events inferred in the MRCA of all peatmosses. Only GO slim terms overrepresented at least in one species are shown. Significantly overrepresented terms (false-discovery rate ≤ 0.05) are shown in black. Non-significant terms are in white. Description for terms overrepresented in at least three species is highlighted.

GO ID	Description	*"F. novo-caledoniae"*	*E. inretortum*	*F. sericeum*	*S. fallax*	*S. lescurii*	*S. subsecundum*	*S. cribrosum*	*S. palustre*	*S. recurvum*
Molecular function ontology										
GO:0003674	molecular_function									
GO:0004518	nuclease activity									
GO:0005215	**transporter activity**									
GO:0005515	**protein binding**									
GO:0016740	**transferase activity**									
GO:0016787	**hydrolase activity**									
GO:0003824	**catalytic activity**									
GO:0005198	**structural molecule activity**									
GO:0030234	enzyme regulator activity									
GO:0005488	**binding**									



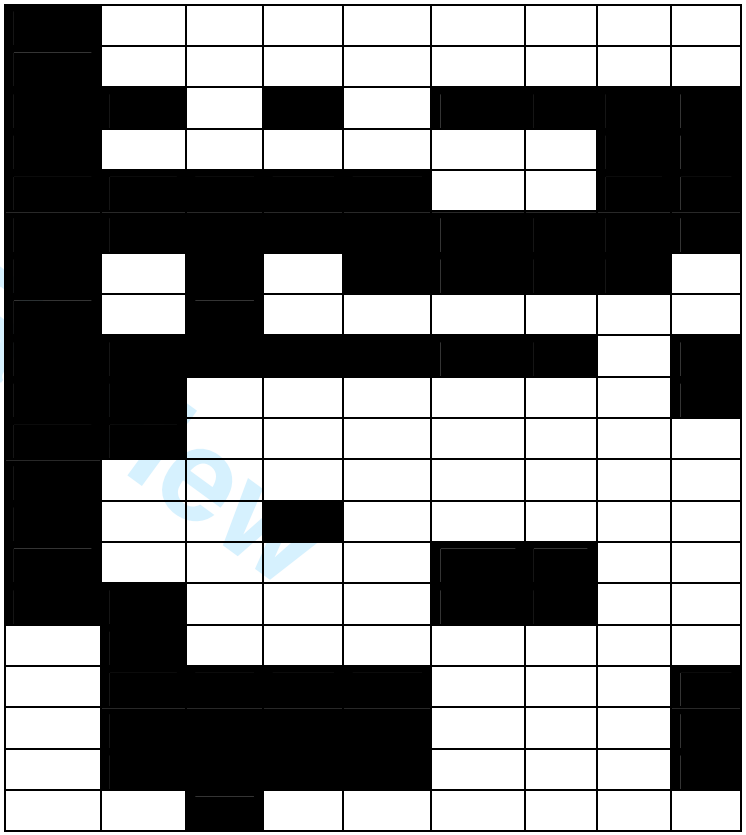
For Peer Review

GO:0003723	RNA binding
GO:0030234	enzyme regulator activity
GO:0003824	catalytic activity



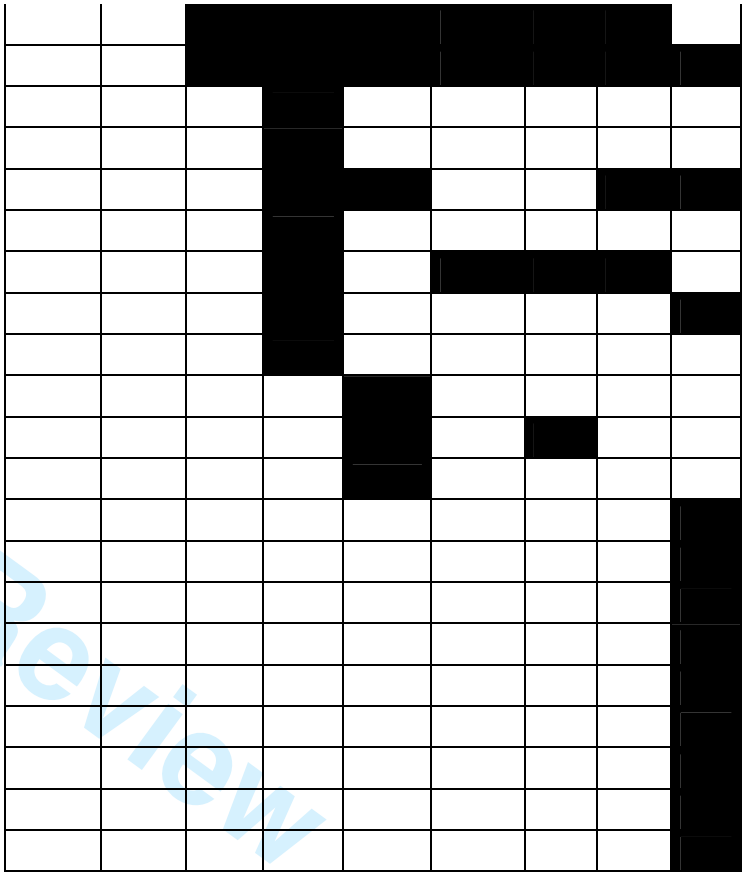
Biological process ontology

GO:0006139	nucleobase-containing compound metabolic process
GO:0006464	cellular protein modification process
GO:0006810	**transport**
GO:0007165	**signal transduction**
GO:0007275	**multicellular organismal development**
GO:0008150	**biological_process**
GO:0008152	**metabolic process**
GO:0008219	cell death
GO:0009628	**response to abiotic stimulus**
GO:0009653	**anatomical structure morphogenesis**
GO:0009791	post-embryonic development
GO:0009908	flower development
GO:0009987	cellular process
GO:0016043	**cellular component organization**
GO:0019725	**cellular homeostasis**
GO:0040007	growth
GO:0006950	**response to stress**
GO:0009605	**response to external stimulus**
GO:0009607	**response to biotic stimulus**
GO:0000003	reproduction



For Peer Review

GO:0006139	**nucleobase-containing compound metabolic process**
GO:0009987	**cellular process**
GO:0019538	protein metabolic process
GO:0015979	photosynthesis
GO:0009719	**response to endogenous stimulus**
GO:0008152	metabolic process
GO:0009058	**biosynthetic process**
GO:0006091	generation of precursor metabolites and energy
GO:0006412	translation
GO:0006629	lipid metabolic process
GO:0009058	biosynthetic process
GO:0009791	post-embryonic development
GO:0009790	embryo development
GO:0009791	post-embryonic development
GO:0009856	pollination
GO:0016043	cellular component organization
GO:0016049	cell growth
GO:0030154	cell differentiation
GO:0040007	growth
GO:0007049	cell cycle
GO:0007154	cell communication

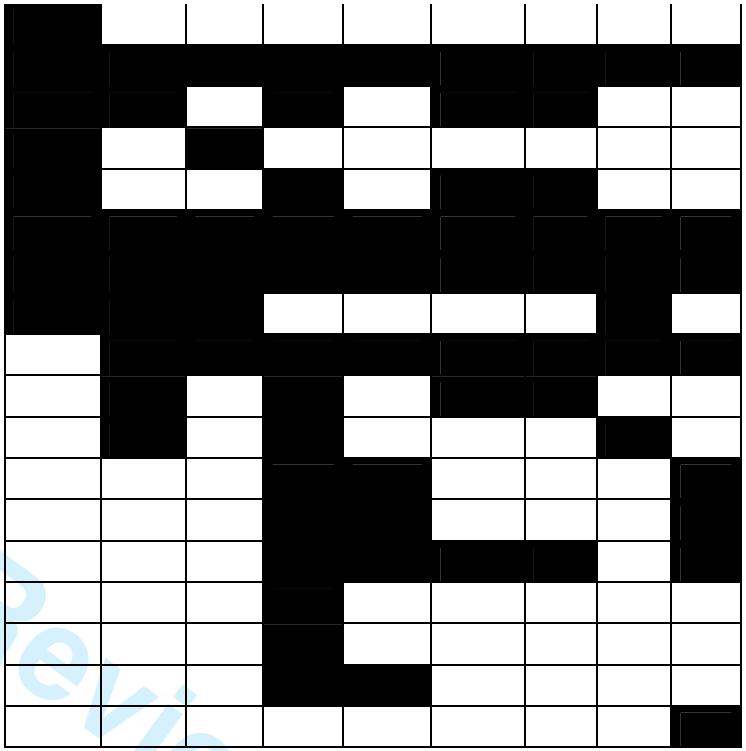


Cellular component ontology

GO:0005575	**cellular_component**
GO:0005622	**intracellular**
GO:0005623	**cell**

For Peer Review

GO:0005635	nuclear envelope
GO:0005737	**cytoplasm**
GO:0005773	**vacuole**
GO:0005777	peroxisome
GO:0005783	**endoplasmic reticulum**
GO:0005829	**cytosol**
GO:0005886	**plasma membrane**
GO:0016020	**membrane**
GO:0005840	**ribosome**
GO:0005730	**nucleolus**
GO:0005618	**cell wall**
GO:0009536	**plastid**
GO:0009579	**thylakoid**
GO:0016020	**membrane**
GO:0005576	extracellular region
GO:0005739	mitochondrion
GO:0005768	endosome
GO:0005794	Golgi apparatus



For Peer Review

Fig. 1

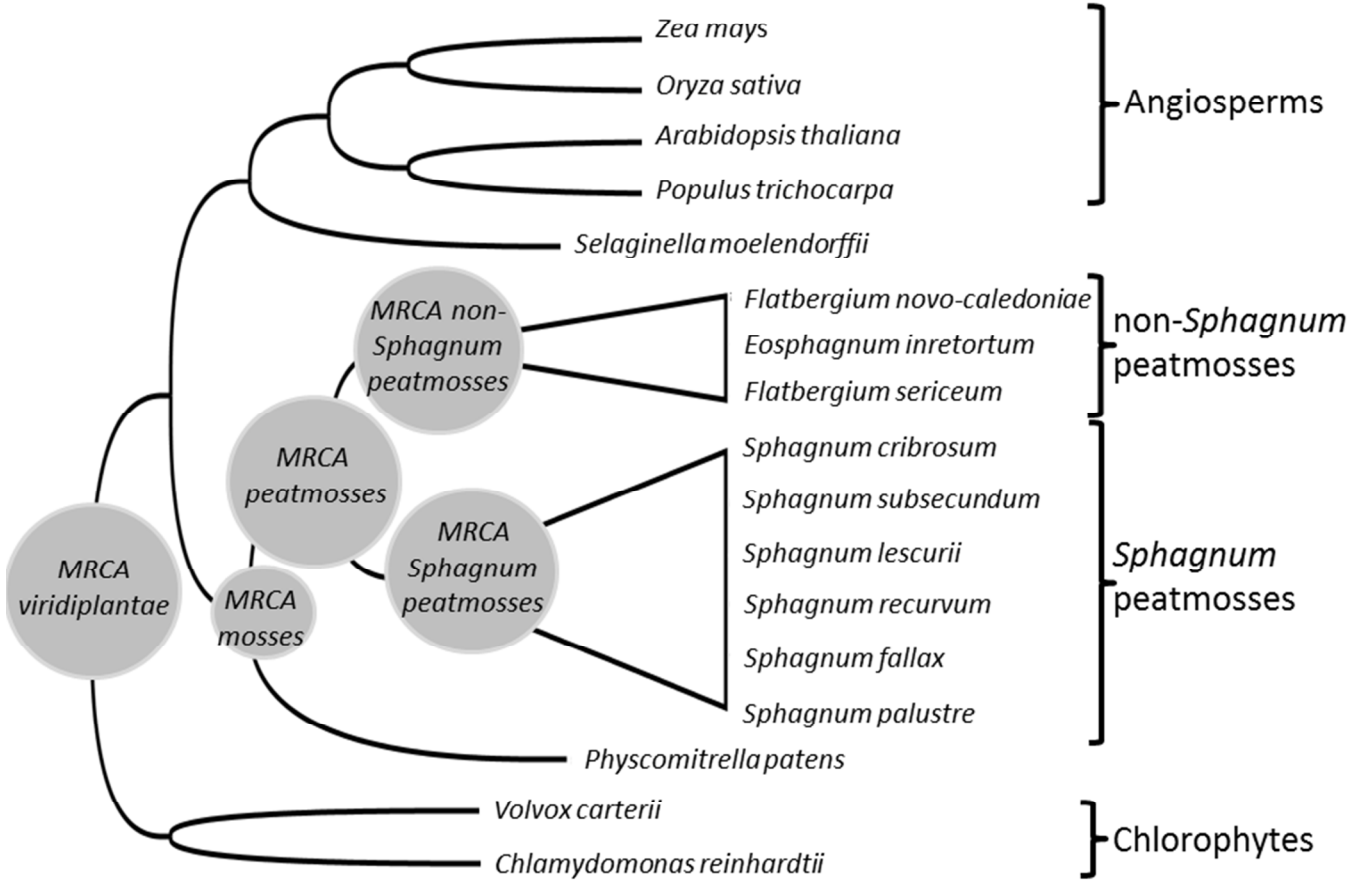

Zea mays
Oryza sativa
Arabidopsis thaliana
Populus trichocarpa
Angiosperms
Selaginella moelendorffii
MRCA non-Sphagnum peatmosses
Flatbergium novo-caledoniae
Eosphagnum inretortum
Flatbergium sericeum
non-Sphagnum peatmosses
MRCA peatmosses
MRCA Sphagnum peatmosses
MRCA viridiplantae
MRCA mosses
Sphagnum cribrosum
Sphagnum subsecundum
Sphagnum lescurii
Sphagnum recurvum
Sphagnum fallax
Sphagnum palustre
Sphagnum peatmosses
Physcomitrella patens
Volvox carterii
Chlamydomonas reinhardtii
Chlorophytes

Fig. 2

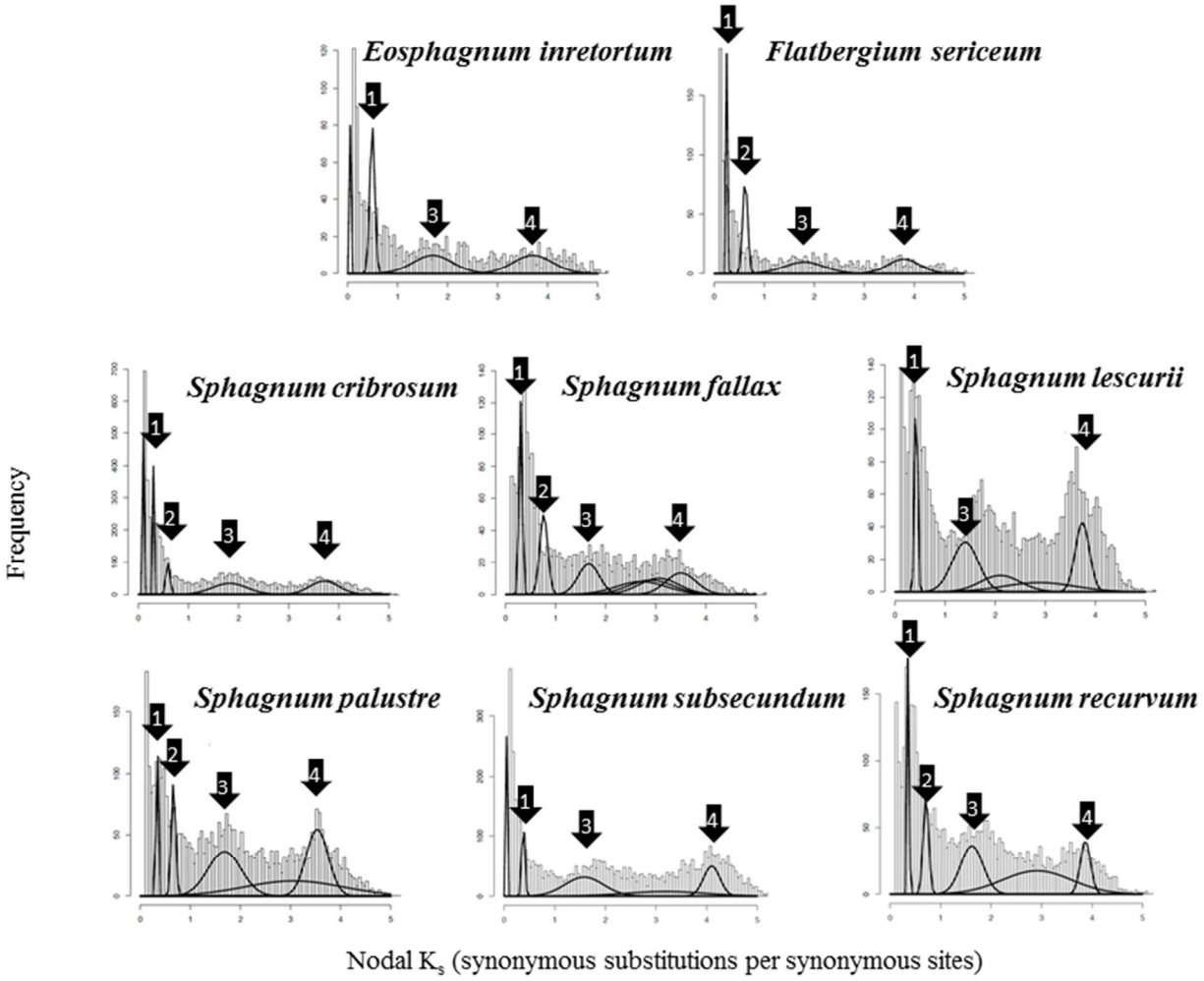

Eosphagnum inretortum
Flatbergium sericeum
Sphagnum cribrosum
Sphagnum fallax
Sphagnum lescurii
Sphagnum palustre
Sphagnum subsecundum
Sphagnum recurvum
Frequency
Nodal K_s (synonymous substitutions per synonymous sites)

Fig 3

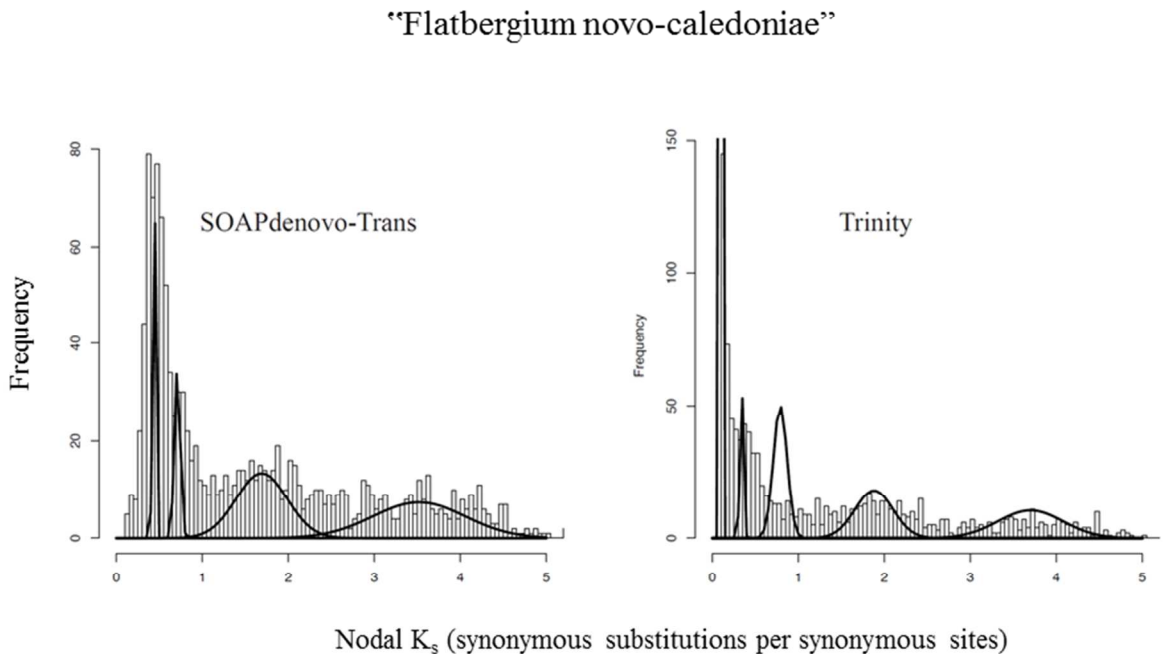
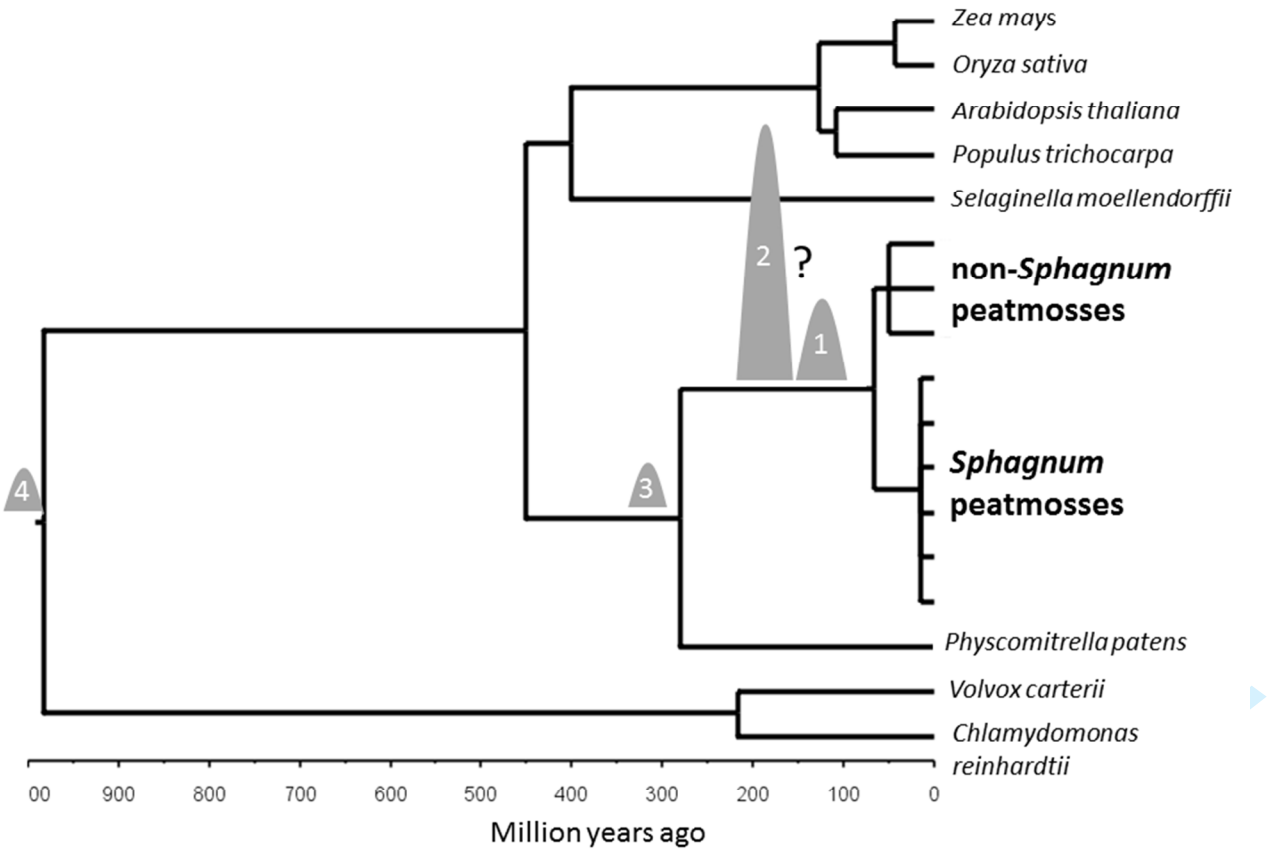

"Flatbergium novo-caledoniae"
SOAPdenovo-Trans
Trinity
Frequency
Frequency
Nodal K_s (synonymous substitutions per synonymous sites)

Fig 4


Zea mays
Oryza sativa
Arabidopsis thaliana
Populus trichocarpa
Selaginella moellendorffii
2
?
1
non-Sphagnum peatmosses
Sphagnum peatmosses
4
3
Physcomitrella patens
Volvox carterii
Chlamydomonas reinhardtii
00
900
800
700
600
500
400
300
200
100
0
Million years ago

For Peer Review